Limits on the Application of Statistical Correlations to Continuous Response Data

Finn Upham

Music and Audio Research Lab, Department of Music and Performing Arts Professions, Steinhardt School of Culture, Education, and Human Development, New York University, USA

finn@nyu.edu

ABSTRACT

How can we compare different listeners' experiences of the same music? For decades, experimenters have collected continuous ratings of tension and emotion to capture the moment-by-moment experiences of music listeners. Over that time, Pearson correlations have routinely been applied to evaluate the similarity between response A and response B, between the time series averages of responses, and between responses and continuous descriptors of the stimulating music. Some researchers have criticized the misapplication and misinterpretation of this class of statistics, but alternatives have not gained wide acceptance. This paper looks critically at the applicability of correlations to continuous responses to music, the assumptions required to estimate their significance, and what is left of the responses when these assumptions are satisfied. This paper also explores an alternative measure of cohesiveness between responses to the same music, and discusses how it can be employed as a measure of reliability and similarity with empirical estimates of significance.

I. INTRODUCTION

Continuous ratings of music perception and experience are common measures of the dynamics of a listener's response. Using some kind of digitally sampled interface, participants report how they perceive or experience the music being presented on scales such as aesthetic experience, tension, and perceived or experienced emotion. Each response forms a time series sampled between 1 and 10 times a second for the duration of the musical stimulus. Although such responses are collected by dozens of researchers around the world, there is little consensus on appropriate techniques for evaluating similarity between responses.

Pearson Product Moment Correlations [PPMC] have been naively applied to these time series since the late 1980s in an attempt to capture the reliability of ratings on repeated tasks [Gregory, 1995]. Correlations have since been employed to compare different participants' responses [Krumhansl, 1996], between sections of responses [Livingstone et al., 2011], and between responses and continuous representations of the music, and to assess legs in responses via cross-correlation [Lucas et al., 2010]. Outside of music cognition work, it is commonly known that correlations cannot be applied blindly to time series data. Schubert in 2002 published an early criticism of the common practice calling out the problem of serial correlation and proposing the practice of analyzing difference data, or reading changes, to reduce the inflation of r-values. Other researchers have attempted to improve matters by using nonparametric correlation measures, such as Spearman [Vines et al., 2006], by downsampling responses to their average Nyquist frequency [Chapin et al., 2010], and by employing autocorrelation models as commonly employed for the analysis of economic time series [Dean and Bailes, 2010].

Despite these warnings and attempts at finding alternatives, researchers have continued to publish analyses of continuous responses using inappropriately applied correlations and estimates of significance. This paper attempts to present in more detail the limits of correlations and the impact of serial correlation in the data, and to deter future abuse of these important classes of calculations.

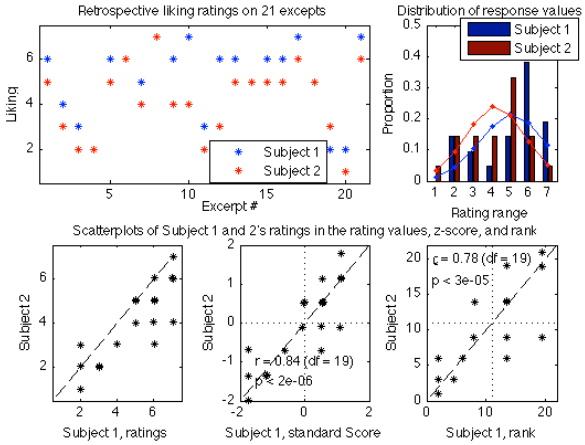

Figure 1. Example of correlation on discrete data: two listeners retrospective ratings of liking on 22 musical excerpts.

II. CORRELATIONS

A correlation is a standardized measure of covariance between two variables [Rodgers and Nicewander, 1988]. Consider the example shown in figure 1 on discrete data: two subjects' retrospective liking ratings for 22 excerpts of music. The top graph to the left shows the values from 1 to 7 which each listener gave to each excerpt. To the right are the distributions of each listener's ratings. Both the bar graph and the estimated normal distribution, $N(\mu_x, \sigma_x)$, capture the fact that on average subject 1 reported lower ratings than subject 2, and this difference is also shown in the left-most scatterplot, as most of the excerpts fall below the diagonal. Correlations discard differences of means and variances to give conveniently interpretable standardized coefficient values. A Pearson product moment correlation between these rating values gives the same result as the Pearson correlation on the data after normalizing each set of ratings to have a unitless distribution with a zero valued mean and a standard deviation of one. The right-most scatterplot of Figure 1 shows the ratings standardized by rank, in which the rating value on each excerpt is replaced by its rank (or in this case its tied rank) from smallest to largest value within each subject's distribution of ratings. This non-linear standardization of values is used to compute the non-parametric Spearman correlation, again discarding units of either variable.

A correlation coefficient calculated on these two sets of liking ratings expresses how closely the listeners' relative

preferences are shared. Though they rarely gave the same rating, it is possible that they agree on which excerpts were worst and which were best. The Pearson Product Moment Correlation coefficient, r, can be calculated with the following equation:

$$r(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{(N-1)\sigma_X\sigma_Y}$$

In this case, X_i and Y_i are the liking ratings given to the i^{th} excerpt by subject 1 and subject 2 respectively. According to the numerator, each excerpt which is rated above average in liking by both subjects contributes positively to the total correlation, as does each excerpt which is considered below average by both. Excerpts on which the subjects disagree (relative to their respective averages) contribute negatively to the total correlation; some of these can be seen in the lower right quadrant of the standard score scatterplot in figure 1. Excerpts rated close to the average liking have little to no impact on the correlation, while those at the extremes of the distributions have considerably more clout. This sensitivity makes the Pearson correlation vulnerable to the effect of outliers; the implications of the values of r, which ranges from -1 to 1, depend on the normality of the distributions being compared, i.e. how well the bell-curve fits over the distribution. Spearman's rho, ρ, is calculated much the same way as the Pearson correlation, using the rank numbers of the variables instead of their measured values. Thus excerpts on which the subjects agree are above their median liking ratings contribute positively to the Spearman correlation coefficient, and so on. The transformation from measured distribution to rank reduces the sensitivity to outliers and does not depend on the assumption of normality. It is a popular non-parametric alternative to the PPMCC, but certainly not the only non-parametric option.

Correlation coefficients by themselves are interesting statistics, analyses of cross-correlation and serial correlation make use of them directly as seen in [Luck et al., 2008]. Correlations are, however, often used in conjunction with a test of significance. Significance tests generally estimate how likely the null hypothesis would yield statistics similar to or more extreme than your empirical data. For correlations, a significance test evaluates how likely the same number of values sampled independently from uncorrelated distributions would result in correlation coefficients of equal or greater value to that of the calculated r or ρ. With some significance threshold such as alpha = 0.01, we agree to consider correlations significant (i.e. presumably repeatable) when the likelihood of the null hypothesis falls below the threshold. Estimating significance is tricky as it depends on our having the right assumptions about this null hypothesis. The default significance estimator in statistics software, the student T estimate for N-2 degrees of freedom, is built on the assumption that the distributions of each variable are normal, i.e. with most values near the mean and steadily fewer values further above and below this central tendency. Another method for estimating the significance of correlation coefficients is to permute the sample values and calculate their correlation a hundred or more times [Hotelling and Pabst, 1936]. The likelihood of the empirical coefficient can then be interpreted with respect to this new distribution. The significance values reported on the correlation graphs (figures 1, 2 and 3) are naïve student T estimates, and they are intended to be interpreted critically. In figure 1, the p values are sufficiently low to ignore this issue.

These significance tests are functions of the number of samples being evaluated: they depend on the assumption that each sample carries independent information about the phenomena under evaluation. The likelihood that coincidence would yield r = 0.4 over 5 samples is much greater than the likelihood that coincidence would yield the same value over 10 or 50 samples. Statistical significance is a measure of the robustness of a relationship between the variables, but it makes no claims about the explanatory or predictive power of one variable on the other.

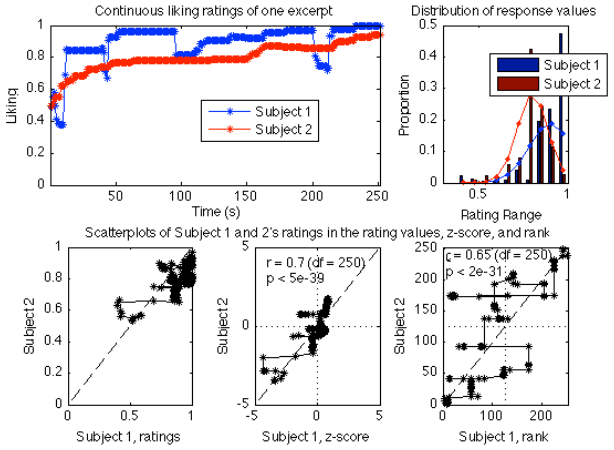

Figure 2. Example of correlation on continuous data: the same two listeners continuous ratings of liking on a four-minute excerpt of music.

Now consider figure 2, the set of graphs depicting the correlation between the same two subjects' continuous ratings of liking on a four-minute musical excerpt. The rating data are sampled every second (1 Hz), a common sample rate on the low side for recent work with continuous rating responses to music. These data demonstrate the three principal challenges of employing correlations on continuous rating responses:

- Serial correlation: these continuous ratings are highly serially correlated, with the value at one point strongly predicting the value of the next. Besides the long plateaus shown in the top-left graph of figure 2, the scatterplots show the data thickly bunched and strung together. Correlation coefficients between time series with high serial correlation are inflated by their respective intra-relatedness [Bartlett, 1935].
- Distribution of values: the values of continuous rating data often fail to be normally distributed. The difference in this case can be seen between the bar graph and the estimated bell curve for each subject's rating values, top-right.
- Independent sampling: given the method of collecting these rating data and the arbitrariness of the sample rate, the number of samples does not represent the amount of independent information [Bartlett, 1935], thus we do not have an easy estimate for the degrees of freedom of any standard significance test.

At 1 Hz, most collections show extremely high average autocorrelation, above 0.85. (top graph). Lowering the sample rate reduces the degree of autocorrelation, but most of these

collections still average around r = 0.6 when sampled every 10 seconds.

III. CORRELATION COMPROMISES

Serial correlation can be assessed by correlating a series with itself delayed by one sample:

$$r(X, X_{lag\ 1}) = \frac{\sum_{i=2}^{N}(X_i - \mu_X)(X_{i-1} - \mu_X)}{(N-2)\sigma_X \sigma_X}$$

One simple method to ameliorate the serial correlation problem is to work instead with the first-order difference of each series, a sequence in which each value reports only how one sample changes from the last in the original data values. First employed on continuous ratings to music by Schubert [Schubert, 2002], this method has been picked up by other researchers in recent years [Dean and Bailes, 2010]. Another possible aid to serial correlation is downsampling, with the assumption that given enough time between samples, new information will have had the chance to influence the rating reported.

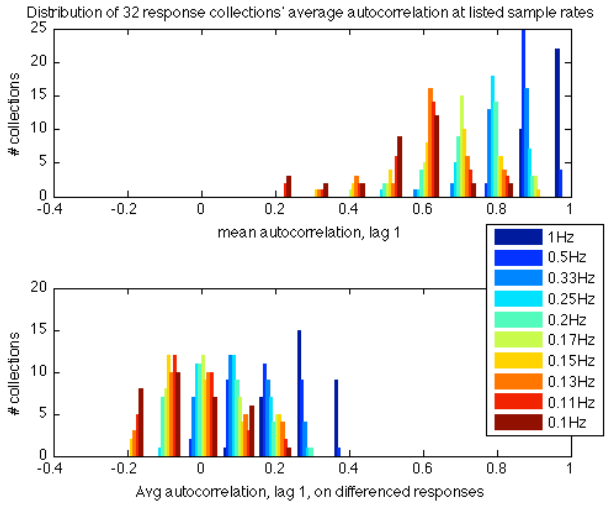

Figure 3. The distributions of average auto-correlation for 32 collections of continuous ratings of emotion to music on the original ratings and their first order difference series at different sample rates.

To test these proposed solutions, Figure 3 shows results from 32 collections of rating responses. These collections are one-dimensional ratings of emotion (valence, arousal, or intensity, experienced and perceived) from five different subject pools (average 27 responses per collection) to 16 different musical stimuli (average length 240 seconds) aggregated from three distinct experiments. The rating ranges have all been normalised to [0,1] and sample rates set to 1 Hz. In the top graph of Figure 3, the distributions of the average autocorrelations of these collections are compared for sample rates going from 1 Hz to 0.1 Hz (once every 10 seconds), and in the bottom graph, the same is evaluated on the first-order difference series of these collections. Differencing the data does dramatically decrease serial correlation, however these collections do not distribute evenly around zero average autocorrelation without also downsampling to 0.167 Hz, or once every 6 seconds. If our primary concern is to eliminate serial correlation from these analyses (rather than compensate using more complicated autoregression models), these data suggest that correlations should be assessed on first-order differenced series which are sampled no faster than once every five seconds. The proportion of negatively autocorrelated collections for very low sample rates suggests a mild oscillation in these series, but the significance of this requires more exploration.

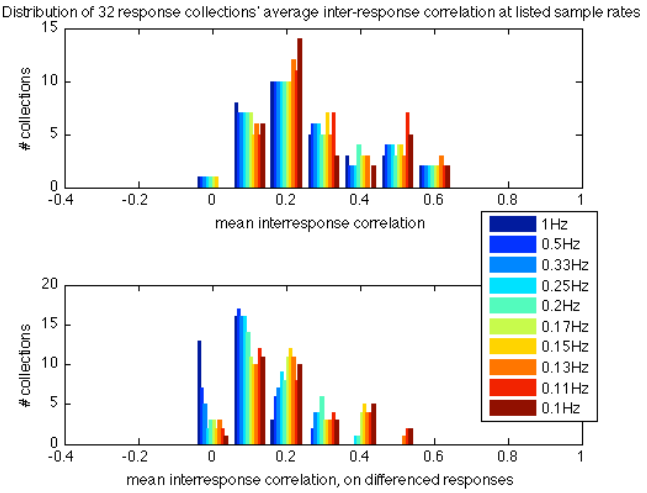

Figure 4. The distributions of average inter-response correlations for 32 collections of continuous ratings of emotion to music on the original ratings and their first order difference series at different sample rates.

Correlations have been used as an argument for the cohesiveness of a collection of responses. Krumhansl's Music Perception article in 1996 may have been the first to report the average inter-response correlation, without acknowledgement of serial correlation on these data. Figure 4 shows how downsampling and differencing the responses in these collections change their average inter-response correlation. Downsampling barely affects this statistic on the original rating data; this underlines the concern that the number of samples in these time series far exceeds the quantity of information they contain. On the other hand, the bottom graph of figure 3 shows the average inter-response correlations of the differenced responses increasing as the sample rate goes down; this happens because each sample is representing a larger time window over which some change of rating may take place. As these correlations do not make use of sequence, finely sampled difference data runs the risk of separating concurrent changes in ratings when participants take different amounts of time to report their reactions to the stimulus. Higher sample rates also have a larger proportion of zero values on these rating data, which can cause further problems for statistical interpretation.

Figure 5 shows the results of applying these reductions to the two responses discussed earlier. The top-left graph shows how the two series are flattened to zero, with variation when the original ratings changed values. Even when downsampled to 0.167 Hz, the distributions are strongly dominated by zero valued data points. The scatterplots show a much less convincing story of the relationship between the two series: the relatively large r value is strongly influenced by one data point, and this advantage is lost in the rank representation of these rating change series.

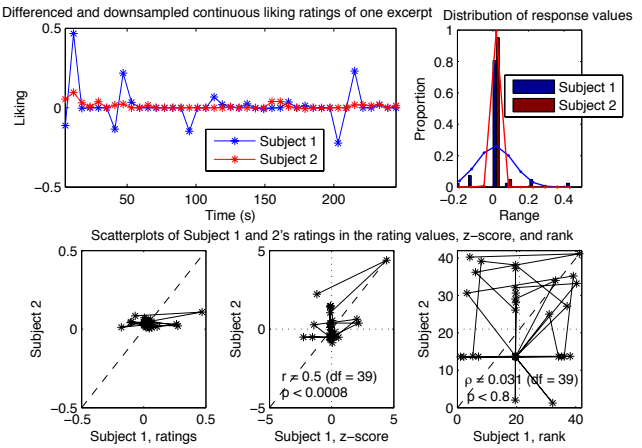

Figure 2. Example of correlation on differenced and downsampled continuous rating data using the same two listeners continuous liking ratings of a four-minute excerpt of classical music.

The goal of reducing serial correlation to zero is somewhat suspect, given the nature of the data. As pointed out by Vines et al., music is serially correlated, at least in its affect [2006], so the low rate of change in continuous ratings may be a reflection of the stimulus rather than marking the upper limit in the temporal sensitivity of the person making the rating.

Looking back at the series plotted in figure 2, it is hard to believe that there is no plausible relationship between these two ratings as implied by the result shown in figure 5. While one subject hardly ever reports any decreases in their liking of the music, the increases in ratings roughly line up. Also, both ratings spend nearly the entire time in the top third of the rating scale. While our eyes easily see such similarities, correlations water them down or throw them away. Considering that regression analysis is also built on the basics of correlations, this suggests that a lot of potentially interesting information has been overlooked by many attempts to analyse these collections of ratings.

IV. ALTERNATIVES TO CORRELATIONS

Alternatives to correlations are numerous, depending on the purpose of the application. The following section concentrates on alternatives to the average inter-response correlation. This is a statistic which has been used to legitimize the cross-sectional average of a collection of responses (the average response time series), depending on the assumption that if the individual responses are highly correlated, their pair wise relationships should be principally driven by the stimulus and thus the average response time series has a good chance of capturing the shared variation of rating values over the many sample points. This has often been reported with estimates of significance, which are (in all likelihood, though the method of estimation is not normally reported) false. Rather than estimate the null hypothesis with shaky assumptions, 32 collections of unrelated responses were assembled by sampling the 32 experimental responses at random and trimming the ends of the responses to fit the shortest in each collection.

One intuitive measure of inter-response similarity is simply their average difference, on the [0,1] rating range, or the average Euclidian distance, normalized for the number of sample points. Figure 6 shows the distribution of the average Euclidean inter-response distances for the real experimental data sets and the random collections in the top graphs: to the left are the original rating series, to the right, the downsampled and differenced series. While the random collections have higher distances between responses, the experimental data sets show a similar range of inter-response distances. The average inter-response Pearson correlation shows similar degrees of overlap between these data, though the statistics reported here are by and large much lower than many of the reported average inter-response correlations in the literature. Applying the same to the differenced data shows less overlap, though the values are still quite low, over all.

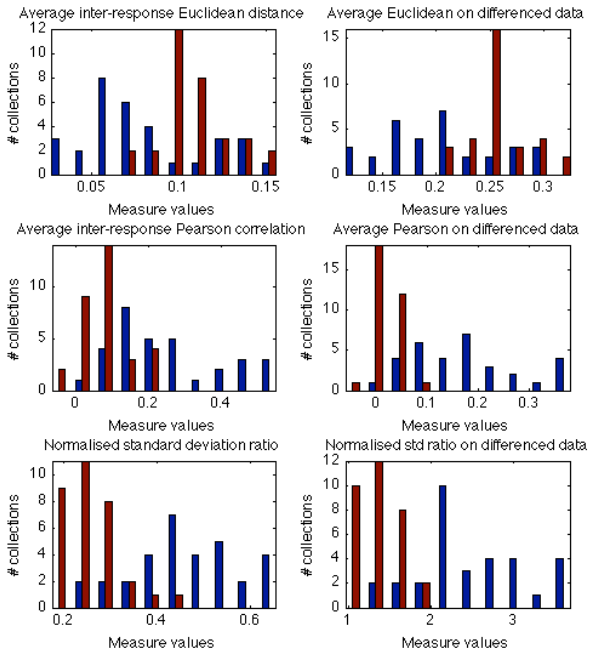

Figure 6. Distributions of coherence statistics on experimental collections of continuous responses and randomly assembled unrelated continuous response collections. The left column shows these measures on rating data, the right column reports the measures as applied to differenced and downsampled responses.

Another alternative measure is inspired by the purpose of the statistic: to validate the average as an aggregation of shared information rather than the accumulation of errors and differences between responses. With some normalization to compensate for the number of responses in each collection, this measure is the ratio of the standard deviation of the average response time series over the average standard deviation of the responses in the collection. By relating the variance of the average rating time series to that of those of the individual responses, we can differentiate between averages that are flat because of disagreement and averages that are flat because the stimulus is not dramatically variable. To capture similarity of contour, the lower right-hand graph of figure 6 shows this measure as applied to the differenced series. The natural logarithm of the number of responses times a constant appears to be a useful method for compensating for

the different sizes of collections, though a theoretical basis for this has not yet been articulated.

To demonstrate the effectiveness of this statistic, the normalized standard deviation ratio of mean- to- individual responses first-order difference series, Figure 7 shows three experimental collections which score high beside three which score within range of the random response collections, their averages plotted in black.

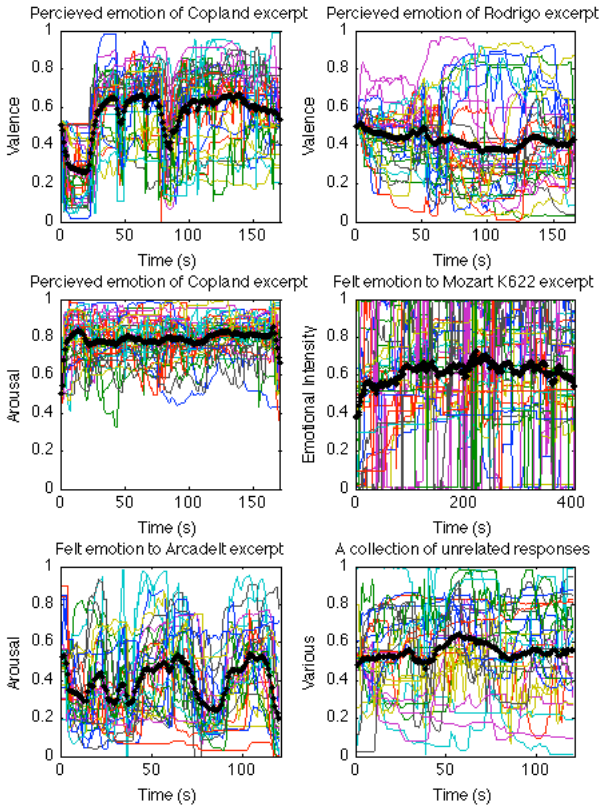

Figure 7. Collections of continuous response separated by the standard deviation ratio test on differenced data. To the left are collections with high cohesiveness, the right are collections which are no more cohesive than the unrelated response collections.

V. CONCLUSION

Continuous rating responses to music do not lend themselves to statistical significance tests of correlations, nor is this class of relatedness always the most relevant, particularly after compensating for serial correlation. There are many contexts in which correlations are used for analysis of continuous ratings of music and musical experience, some of which are not seriously hampered by the issues discussed above. Cross correlation with descriptions of the stimuli, for example, can be employed so long as some care is taken to establish reasonable limits on confidence. Very high correlation coefficients (say above 0.9, or 0.5 for differenced data) can probably be trusted as a strong sign of relatedness, even with high serial correlation. But estimates of significance for these coefficients should only be included with explanations of how the three main issues, serial correlation, non-parametric distributions, and independent sampling, have been addressed.

It is worth exploring measures of similarity or relatedness that are closer to our intuitions on these rating data. Many publications have repeated the same analytic mistakes out of convenience while potentially important information has gone unreported. While many others have handled some of these issues, broadening awareness of the challenges of time series in music cognition is the only way to find better methods. Employing data from many experiments makes this kind of comparative analysis possible, and I hope that others opt to share published experimental data to improve the analytic power of future methodological research.

ACKNOWLEDGMENT

My thanks to Stephen McAdams and Mark Korhonen for sharing their continuous rating data; to NSERC and SSHERC, funding bodies which made the collection of these data possible, and the many people involved in the BSO and CARS experiments; to Stephen McAdams again for supervising the beginning of this research as part of my master's thesis, and to my current advisor, Mary Farbood, and the Steinhardt School at NYU for supporting its continuation.

REFERENCES

Bartlett, M. (1935). Some aspects of the time-correlation problem in regard to tests of significance. *Journal of the Royal Statistical Society*, 98(3):536–543.

Chapin, H., Jantzen, K., Scott Kelso, J., Steinberg, F., Large, E., and Rodriguez-Fornells, A. (2010). Dynamic emotional and neural responses to music depend on performance expression and listener experience. PloS one, 5(12):169–200.

Dean, R. and Bailes, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception of music. Empirical Musicology Review, 5(4):152–175.

Gregory, D. (1995). The continuous response digital interface: an analysis of reliability measures. Psychomusicology, 14:197–208.

Hotelling, H. and Pabst, M. (1936). Rank correlation and tests of significance involving no assumption of normality. The Annals of Mathematical Statistics, 7(1):29–43.

Krumhansl, C. L. (1996). A perceptual analysis of mozart's piano sonata k. 282: Segmentation, tension, and musical ideas. Music Perception, 13(3):401–432.

Livingstone, S., Palmer, C., and Schubert, E. (2011). Emotional response to musical repetition. Emotion, Epub.

Lucas, B., Schubert, E., and Halpern, A. (2010). Perception of emotion in sounded and imagined music. Music Perception, 27(5):399–412.

Luck, G., Troiviainen, P., Erkkilä, J., Lartillot, O., Riikkilä, K., Mäkelä, A., Pyhäluoto, K., Raine, H., Varkila, L., and Värri, J. (2008). Modelling the relationships between emotional responses to, and musical content of, music therapy improvisations. Psychology of Music, 36(1):25–45.

Rodgers, J. and Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1):59–66.

Schubert, E. (2002). Correlation analysis of continuous emotional response: Correcting for the effects of serial correlation. Musicae Scientiae, Special Issue 2001-2002:213–236.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. Music Perception, 21(4):561–585.

Vines, B., Krumhansl, C., Wanderley, M., and Levitin, D. (2006). Cross-modal interactions in the perception of musical performance. Cognition, 101(1):80–113.

Wöllner, C. and Auhagen, W. (2008). Perceiving conductors' expressive gestures from different visual perspectives. an exploratory continuous response study. Music Perception, 26(2):129–143.